Published online 16 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D363–D366***
doi:10.1093/nar/gkl970

euHCVdb: the European hepatitis C virus database

Christophe Combet*, Nicolas Garnier, Céline Charavay, Delphine Grando, Daniel Crisan, Julien Lopez, Alexandre Dehne-Garcia, Christophe Geourjon, Emmanuel Bettler, Chantal Hulo[1], Philippe Le Mercier[1], Ralf Bartenschlager[2], Helmut Diepolder[3], Darius Moradpour[4], Jean-Michel Pawlotsky[5], Charles M. Rice[6], Christian Trépo[7], François Penin and Gilbert Deléage

Institut de Biologie et Chimie des Protéines (UMR5086), CNRS, University of Lyon 1, IFR 128 BioSciences Lyon-Gerland, France, [1]Swiss Institute of Bioinformatics, Geneva, Switzerland, [2]Department for Molecular Virology, University of Heidelberg, Germany, [3]Institute for Immunology, University of Munich, Germany, [4]Division of Gastroenterology and Hepatology, Centre Hospitalier Universitaire Vaudois, University of Lausanne, Switzerland, [5]Department of Virology, INSERM U635, Henri Mondor Hospital, University of Paris XII, Creteil, France, [6]Center for the Study of Hepatitis C, Laboratory of Virology and Infectious Disease, The Rockefeller University, New York, NY and [7]INSERM U271, Lyon, France

Received August 10, 2006; Accepted October 20, 2006

ABSTRACT

The hepatitis C virus (HCV) genome shows remarkable sequence variability, leading to the classification of at least six major genotypes, numerous subtypes and a myriad of quasispecies within a given host. A database allowing researchers to investigate the genetic and structural variability of all available HCV sequences is an essential tool for studies on the molecular virology and pathogenesis of hepatitis C as well as drug design and vaccine development. We describe here the European Hepatitis C Virus Database (euHCVdb, http://euhcvdb.ibcp.fr), a collection of computer-annotated sequences based on reference genomes. The annotations include genome mapping of sequences, use of recommended nomenclature, subtyping as well as three-dimensional (3D) molecular models of proteins. A WWW interface has been developed to facilitate database searches and the export of data for sequence and structure analyses. As part of an international collaborative effort with the US and Japanese databases, the European HCV Database (euHCVdb) is mainly dedicated to HCV protein sequences, 3D structures and functional analyses.

INTRODUCTION

Hepatitis C virus (HCV) infection is a major cause of chronic hepatitis, liver cirrhosis and hepatocellular carcinoma worldwide. The HCV genome is ~9600 nt in length and carries a single, long open reading frame (ORF) flanked by 5′ and 3′ non-translated regions. The ORF encodes a polyprotein of ~3000 amino acids that is processed by cellular and viral proteases to yield at least 10 mature proteins: C, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B (1,2). The sequence diversity among HCV genomes leads to the definition of a large number of subtypes distributed into six genotypes (3). In addition, HCV exists within its hosts as a pool of genetically distinct but closely related variants referred to as quasispecies (4). It is now well established that the genotype is a predictive factor of the response to interferon-α therapy (5). Consequently, intensive sequence analyses of HCV genomes are currently conducted, and >40 000 nt sequences have been deposited to date into the DDBJ/EMBL/GenBank databases. In order to manage such large and growing collections of sequences, to facilitate sequence analysis and drug and vaccine design, several specialized databases have been developed (6). The database team members are involved in a network of experts for HCV nomenclature definition and harmonization (3). We present here the European HCV Database (euHCVdb). This database combines computer-annotated HCV sequences with protein three-dimensional (3D) models and it is linked to numerous sequence and structure analysis tools on dedicated websites. The euHCVdb is mainly oriented towards the structural biology of HCV, including protein sequence, structure and function analyses (2).

DATABASE CONSTRUCTION

The euHCVdb is an extension of the French HCV Database developed in 1999 (7). It is updated on a monthly basis from

*To whom correspondence should be addressed. Tel: +33 4 37 65 29 47; Fax: +33 4 72 72 26 04; Email: c.combet@ibcp.fr

Members of the euHCVdb Scientific Board: Ralf Bartenschlager, Helmut Diepolder, Darius Moradpour, Jean-Michel Pawlotsky, Charles M. Rice, Christian Trépo and François Penin

the EMBL nucleotide sequence database (8) and maintained in the PostgreSQL relational database management system (RDMS). The programs for parsing the EMBL database flat files, annotating HCV entries, filling up and querying the database use SQL and Java. Great effort has been devoted to developing a fully automated annotation procedure using a reference set of 26 well-characterized complete HCV genomes representing 18 subtypes (3). The molecular models based protein homology are automatically computed and stored in a separate database by a protein molecular model database management system written in Python and called Modeome3D (N. Garnier, C. Combet, C. Geourjon, G. Deléage and E. Bettler, manuscript in preparation). The building procedure starts by the parsing of the EMBL database (*vrl* section) flat text file. The entries corresponding to the organism HCV are loaded into the RDMS. The second step of the procedure is the sequence annotation, which results in the creation of the euHCVdb and the euHCVdb3D databases by adding annotations to the EMBL data (i.e. genome mapping, protein features, genotype/subtype or 3D models). The sets of pre-computed multiple sequence alignments of complete proteins or reference genomes and 3D models are also updated monthly.

DATABASE CONTENTS

The format of euHCVdb is an extension of the EMBL database format. Thus, the primary accession number, the creation date, the bibliographic references, the *source* feature and the sequence of each EMBL database entry are conserved in the euHCVdb entries. The euHCVdb entry identifiers are built from the EMBL primary accession number to avoid any change in references for sequences. A text search on the EMBL *source* feature is performed to retrieve subtype and isolate name of the deposited sequence. They are stored in dedicated qualifiers in the euHCVdb entries. The sequence of the entry is then retrieved and used to run a FASTA (9) similarity search against the sequences of the reference genomes. The annotations of the most similar genome are retrieved to map regions and features on the current sequence. The current entry is then subtyped by using two sets of reference nucleotide sequences as described in (3). These reference sequences cover two subgenomic regions, one in C/E1 and one in NS5B proteins, and are stored in two sequence databanks. A FASTA search is run for each sequence against each databank and the provisional genotype field of the entry is filled only if the following requirements are satisfied: sufficient coverage of the alignment, similarity level and consistency of genotype between the two subgenomic regions. The annotation procedure ends by looking for protein 3D structural templates to build 3D models of the HCV proteins by homology modeling. For each sequence, a similarity search is run against a selected set of HCV protein structures deposited in the Protein Data Bank (PDB) (10). When acceptable alignments are found, the 3D models are computed using the Geno3D program for automatic comparative homology modeling of proteins (11). A product feature (*prod_ft* qualifier) indicated by *model3d*, which includes the template PDB code, is added to the corresponding *mat_peptide* feature of the entry and the 3D model is stored in the euHCVdb3D. Each euHCVdb entry offers external cross-references to the NCBI taxonomy, the UniProt knowledgebase (UniProtKB) (12) and to EMBL database, as well as internal cross-references to the reference genome used for annotation and to the euHCVdb3D molecular models. The euHCVdb database is also cross-referenced by the UniProtKB/Swiss-Prot (12) and LANL HCV (13) databases and by the NCBI LinkOut system. All these cross-references allow rapid collection of additional information when required.

The automatic annotation procedure ensures standardized nomenclature for all entries across the database and builds a description of the HCV genomic regions and proteins that are included in the entry. This procedure also provides bibliographic references, cross-references to external databases, genotype, well-characterized sites (e.g. hypervariable region 1, HVR1) or domains (e.g. NS3 helicase), the source of the sequence (e.g. isolate), and structural data as 3D protein models. The protein annotations are done in close collaboration with the Swiss-Prot group of the Swiss Institute of Bioinformatics. The format and the controlled vocabulary of the UniProtKB/Swiss-Prot database are used in euHCVdb at the level of the *prod_ft*.

WEB INTERFACE

The euHCVdb is accessible through a website (http://euhcvdb.ibcp.fr) (Figure 1). It is divided into static and dynamic parts. The static part allows the user to access the description of genomic regions or proteins by clicking on pictures (Figure 1A). Pre-computed multiple sequence alignments of reference genomes or complete protein sequences can be viewed and edited with the EditAlignment java Applet developed in our team (Figure 1B). When the experimental 3D structures of the molecules are known, links to PDB files are available that allow users to view and analyze the structure using the Jmol applet. Users can also find recommended nomenclature, the list of the reference genomes and links to other resources in the static part.

In the dynamic part, a query system (not shown) allows the building of dynamic sets of sequences or 3D models according to user-defined criteria (>30 different criteria can be defined) and selected by the user through the query interface, e.g. 'extract all the 3D models of the NS3 protease of confirmed genotype 3, 4 and 5'. The results (Figure 1C) are displayed in a table where each row corresponds to an entry or a genomic region/protein that is described by a small set of identifiers and characteristics such as accession number, protein name, genotype, isolate, length and description. Links to other HCV databases are also mentioned in this table. A hypertext link on the accession number of each entry allows display of the complete entry data in the *Entry details* page (Figure 1D). This page contains hyperlinks to external resources such as EMBL database, to retrieve partial or complete nucleotide or protein sequences of the entry or to the Jmol Web page to analyze pre-computed 3D protein models when available (Figure 1E). The results in Figure 1C can be ordered by ascending/descending values of each column. Above the table, a form allows the user to change the number of results by page or to switch to a given page or accession number. A toolbox is also available to export data (e.g. sequences) to other resources. Data to be exported depend

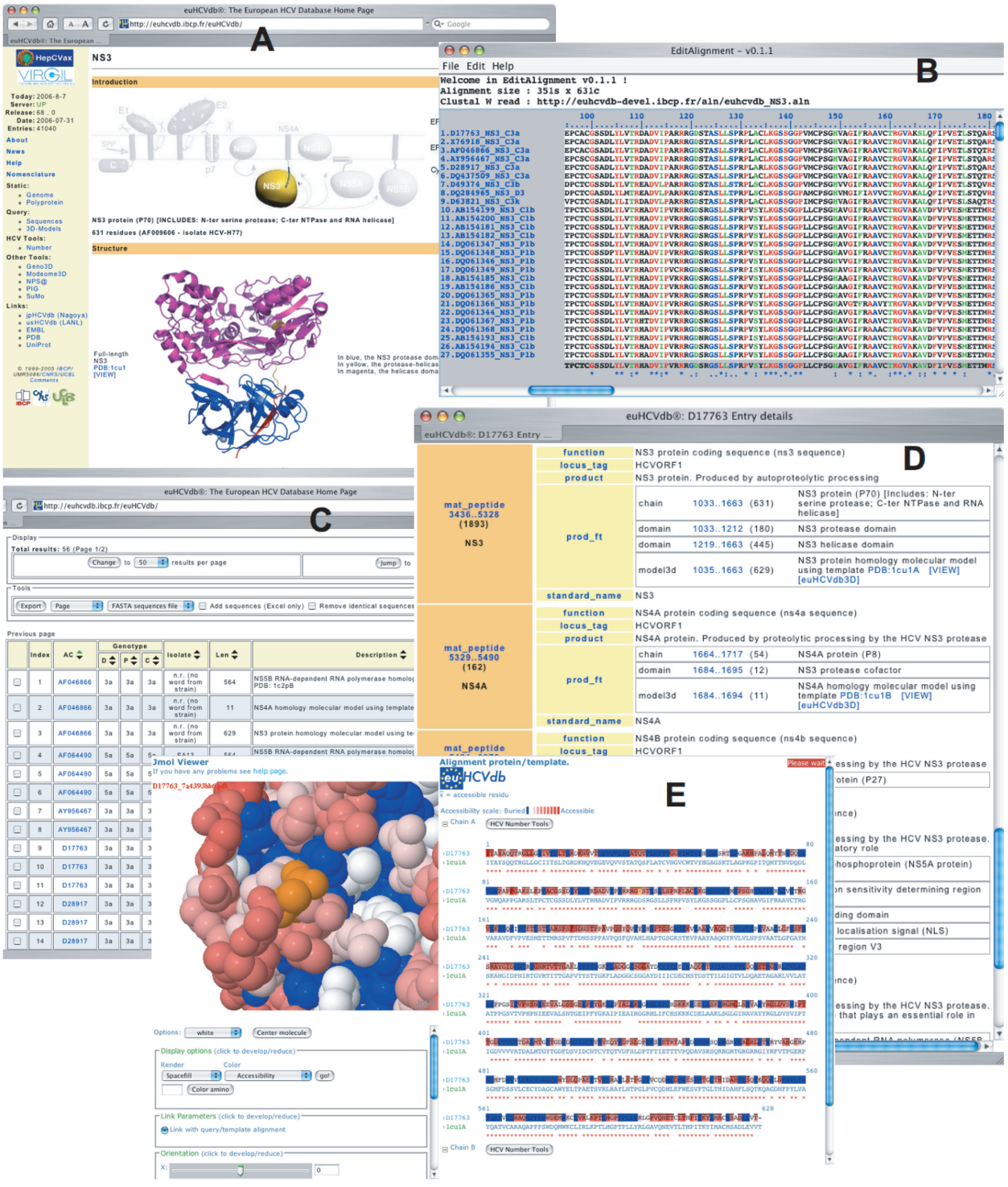

Figure 1. An overview of euHCVdb. (**A**) A static page describing known data for NS3 protein. (**B**) The multiple sequence alignment of NS3 protein from (**A**) viewed in the EditAlignment applet. (**C**) A query result page with the display and tools boxes, and list of hits. (**D**) An entry details page, with data for NS3 and links to the euHCVdb3D models. (**E**) A molecular model viewed through the euHCVdb3D interface. The Jmol applet is linked with the alignment between the modeled and template PDB sequences.

on the *Sequence type* selected in the query form. Available sequence types are the full-length nucleotide sequence or the corresponding polyprotein, a genomic region or a protein, or a sequence corresponding to a polynucleotide or protein feature. The toolbox allows the export of all/page/(un)selected results as a text file of entries, a Pearson sequence file or a tab-delimited text file. Sequences can be transferred to the NPS@ server (14) which is an integrated sequence analysis Web server with a simple interface that provides the access to 46 programs for sequence analyses (e.g. BLAST and

CLUSTAL W) and 12 biological databases. With the support of this server, the user can, for example, obtain a multiple sequence alignment of the full-length NS3 protein sequences to analyze the amino acid variability at each position of the alignment with the residue repertoire tool (F. Dorkeld, F. Penin, G. Deléage and C. Combet, unpublished data). In addition, the 3D models of different variants can be extracted from the database and further analyzed using the PIG Web server for protein structure analysis (N. Garnier, G. Deléage and E. Bettler, manuscript in preparation). Selected 3D models can be superimposed to generate a multiple structural alignment. The user can interactively analyze the fitted models by using the Jmol applet (Figure 1E) that is dynamically linked to the corresponding sequence alignment. In this way, one can identify the differences between the variants. Such analysis is helpful for a better understanding of structure–function relationships, which is of high relevance, e.g. for understanding drug resistance. The user can also visualize the residue conservation at the model level, obtain a list of accessible residues, or detect ligand-binding or active sites using the SuMo Web server (15).

DATABASE STATISTICS

The database has been running since March 2005. The current release number 71 (October 2006) comprises 43 124 entries and 646 protein 3D models representing 914 protein chains for a total of 383 443 residues and 2 930 254 atoms. The database currently receives ~200 queries a day.

CONCLUSIONS

The relational database of annotated HCV sequences and protein 3D models we have developed focuses on HCV protein sequence, structure and function analyses. The automatic annotation process used to generate euHCVdb guarantees the consistency of annotations across the database, which allows efficient keyword searches and comprehensive sequence and structure analyses. The euHCVdb website permits dynamic queries through a very simple interface, and query results can be further analyzed with a set of bioinformatics programs available in the NPS@ and PIG Web servers. By providing the ability to conduct complex searches and analyses, euHCVdb is a powerful tool for researchers working in the HCV field. Moreover, we are working in an international collaborative effort with the US and Japanese HCV databases to improve the use of harmonized HCV data and nomenclature, and making an effort to be as complementary as possible with the goal of providing helpful and efficient tools for HCV researchers around the world.

ACKNOWLEDGEMENTS

We would like to thank Drs Carla Kuiken and Masashi Mizokami for the fruitful collaboration between the HCV databases as well as Dr Peter Simmonds and all members of the network for HCV nomenclature for stimulating discussions. The European HCV Database is funded by the European Commission (HepCVax Cluster, FP5 grant QLK2-2002-01329 and VIRGIL Network of Excellence, FP6 grant LSHM-CT-2004-503359) and Agence Nationale de Recherches sur le SIDA et les Hépatites Virales (ANRS). Nicolas Garnier is the recipient of an Association Française contre les Myopathies (AFM) doctoral fellowship. Funding to pay the Open Access publication charges for this article was provided by European Commission (VIRGIL Network of Excellence, grant LSHM-CT-2004-503359).

Conflict of interest statement. None declared.

REFERENCES

1. Lindenbach,B.D. and Rice,C.M. (2005) Unravelling hepatitis C virus replication from genome to function. *Nature*, **436**, 933–938.
2. Penin,F., Dubuisson,J., Rey,F.A., Moradpour,D. and Pawlotsky,J.M. (2004) Structural biology of hepatitis C virus. *Hepatology*, **39**, 5–19.
3. Simmonds,P., Bukh,J., Combet,C., Deleage,G., Enomoto,N., Feinstone,S., Halfon,P., Inchauspe,G., Kuiken,C., Maertens,G. *et al.* (2005) Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*, **42**, 962–973.
4. Weiner,A.J., Christopherson,C., Hall,J.E., Bonino,F., Saracco,G., Brunetto,M.R., Crawford,K., Marion,C.D., Crawford,K.A., Venkatakrishna,S. *et al.* (1991) Sequence variation in hepatitis C viral isolates. *J. Hepatol*, **13** (Suppl. 4), S6–S14.
5. Pawlotsky,J.M. (2003) Hepatitis C virus genetic variability: pathogenic and clinical implications. *Clin. Liver Dis.*, **7**, 45–66.
6. Kuiken,C., Mizokami,M., Deleage,G., Yusim,K., Penin,F., Shin,I.T., Charavay,C., Tao,N., Crisan,D., Grando,D. *et al.* (2006) Hepatitis C databases, principles and utility to researchers. *Hepatology*, **43**, 1157–1165.
7. Combet,C., Penin,F., Geourjon,C. and Deleage,G. (2004) HCVDB: hepatitis C virus sequences database. *Appl. Bioinformatics*, **3**, 237–240.
8. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P., van den Broek,A. *et al.* (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.*, **34**, D10–D15.
9. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
10. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
11. Combet,C., Jambon,M., Deleage,G. and Geourjon,C. (2002) Geno3D: automatic comparative molecular modelling of protein. *Bioinformatics*, **18**, 213–214.
12. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
13. Kuiken,C., Yusim,K., Boykin,L. and Richardson,R. (2005) The Los Alamos hepatitis C sequence database. *Bioinformatics*, **21**, 379–384.
14. Combet,C., Blanchet,C., Geourjon,C. and Deleage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.
15. Jambon,M., Andrieu,O., Combet,C., Deleage,G., Delfaud,F. and Geourjon,C. (2005) The SuMo server: 3D search for protein functional sites. *Bioinformatics*, **21**, 3929–3930.